이동통신 환경 하에서의 서버 기반 음성 인식을 위한 음성 부호화 기법

이길호, 윤재삼, 오유리, 김홍국
광주과학기술원 정보통신공학과

A Speech Coder for Server-Based Speech Recognition in Mobile Communication

Gil Ho Lee, Jae Sam Yoon, Yoo Rhee Oh, and Hong Kook Kim
Gwangju Institute of Science and Technology (GIST)
E-mail : {ghlee, jsyoon, yroh, hongkook}@gist.ac.kr

요약

본 논문의 목적은 이동통신 환경 하에서 음성 인식과 음성 부호화를 성능의 저하 없이 동시에 수행하기 위한 기법을 개발하는 것에 있다. 이를 구현하기 위해 통신상에서 전송되는 음성 특징 파라미터는 기존 음성 부호화기의 LPC 대신 음성 인식 파라미터인 MFCC를 사용하였다. 따라서 음성 인식 성능은 향상된다. 하지만 음성 재생을 위해 MFCC를 LPC로 변환하는 과정에서 오차가 발생하여 전송되는 bit 수에 비해 만족할만한 음질을 얻을 수 없다. 따라서 이 오차를 보상하여야 하며 이를 위한 변수를 추가하여 음질을 개선시켰다. 그 결과 음질과 음성 인식에서 안정된 성능을 보이는 음성 부호화기를 개발하였다.

1. 서론

이동통신 (또는 무선통신) 환경 하에서의 Automatic speech recognition (ASR) 방식은 특징 파라미터의 추출과 인식을 어디서 담당하는 가에 따라 크게 세 가지로 나눌 수 있다. 다시 말해, 음성 인식을 단말기에서 하는지 혹은 통신망에 연결된 서버에서 하는 지에 따라 음성 인식 시스템을 분류할 수 있다. 즉, 단말기 기반 방식 (client-based system 혹은 embedded system), 서버 기반 방식 (server-based system 혹은 bitstream-based system), 그리고 이들 두 가지 방식을 혼합한 단말기/서버 기반 방식 (client/server-based system 혹은 distributed speech recognition (DSR) system)이 있다 [1].

본 논문에서는 위와 세 가지 방식중 서버 기반 방식의 ASR 관점에서 양질의 성능을 보이는 음성 부호화기 개발에 목적을 두었다. spectral envelope을 구하기 위한 기존 방식은 linear prediction coefficients (LPC)와 같은 파라미터를 통한 estimation이었다. 하지만 본 논문에서 제안하고 있는 음성 부호화기는 spectral envelope을 mel-frequency cepstral coefficients (MFCC)를 통해서 구하고 있다. 따라서 기존 음성 부호화기를 이용한 서버 기반의 ASR 시스템보다 음성 인식의 관점에서 보다 나은 성능을 보이게 되는 것이다.

2. 음성 인식을 위한 음성 부호화기의 개요

본 논문에서는 단순한 음성 부호화기가 아닌 음성 인식기에서도 충분한 성능을 보이는 음성 부호화기를 제안하고 있다. 제안된 음성 부호화기는 DSR의 개념과 음성 부호화의 특징 분석, 음성 부호의 양자화를 조합하여 구현하였다.

서버 기반 방식에서의 음성 인식률을 유지하기 위해 spectral envelope을 MFCC로부터 구하는 code-excited linear prediction (CELP) 음성 부호화기를 설계했다. 기존 CELP 음성 부호화기는 spectral envelope을 LPC로부터 구하고 그 LPC를 양자화하여 전송하게 된다. 그러나 제안된 음성 부호화기에서는 음성 인식율을 고려하여 음성 특징 파라미터로 MFCC를 구하여 양자화한 후 전송한다. 따라서 음성 재생을 위해 MFCC를 LPC로 변환하는 과정이 필요하다. 제안된 음성 부호화기는 그림 1에 나타나있다.

제안된 음성 부호화기는 ITU-T Recommendation G.729 [2]를 기본으로 하여 구현하였다. 즉, 10ms 마다 frame을 구성하며 각 frame은 long-term prediction과 excitation 모델링을 위해 두 개의 subframe으로 나뉘게 된다. 제안된 음성 부호화기가 G.729와 다른 점은 MFCC extraction, MFCC 양자화 그리고 MFCC를 LPC로 바꾸는 conversion 과정이 추가된 것이다. 제안된 음성 부호화기의 전체 bit 할당 정보는 표 1에 나타내었다.

그림 2는 MFCC extraction 과정을 보여주고 있다. 음성 신호는 140 Hz의 cut-off 주파수를 갖는 high-pass 필터를 거쳐 2로 나누어지는 preprocessing을 하게 된다. 다음으로 G.729에서 사용하는 비대칭 윈도우를 통과하게 된다. 그 후 256개의 sample로 zero padding을 한 후 256 point의 FFT를 하여 magnitude spectrum이 계산되어진다. 이 magnitude spectrum이 23개의 triangular mel-filterbank와 로그 스케일, discrete cosine transform의 단계를 거쳐 MFCC가 된다. 이 23개의 MFCC 중 13개가 음성 인식에 사용된다 [4].

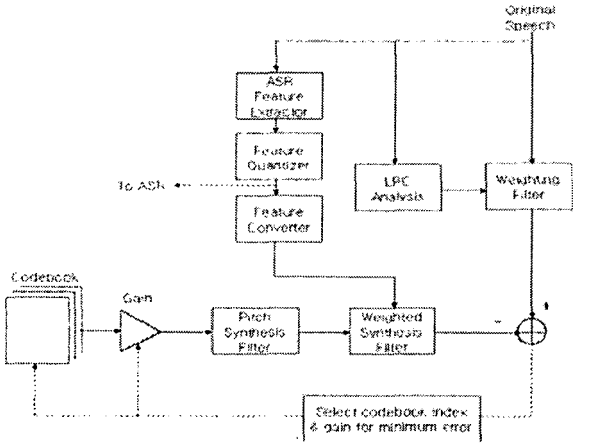

그림 1. 제안된 음성 부호화기

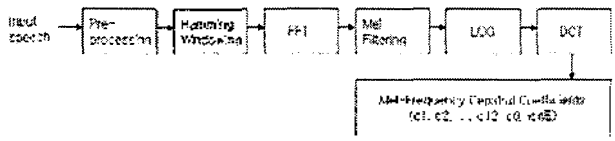

그림 2. MFCC extraction 과정

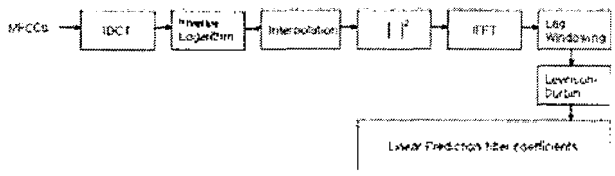

그림 3. MFCC를 LPC로 변환하는 과정

Parameter	Subframe		Frame
	1	2	
MFCC	-		44
Adaptive codebook index	8	5	13
Pitch parity	1	-	1
Fixed codebook index	13	13	26
Fixed codebook sign	4	4	8
Codebook gain	7	7	14
Total			106

표 1. 제안된 음성 부호화기의 bit 할당 정보

MFCC를 LPC로 변환하는 과정은 그림 3에 나타나있다. 먼저 13개의 MFCC는 zero padding을 거쳐 23개의 MFCC로 만들어지고 이 값들은 역 DCT와 역 log scale을 거치게 된다. 다음으로 23개의 값들은 256개의 값으로 선형 보간을 하고 각 값들을 제곱하여 역 FFT 과정을 거치면 spectrum을 구할 수 있다. 이 spectrum을 통해 최종적으로 LPC를 구할 수 있다.

표 2는 원래 음성으로부터 구한 LPC와 G.729의 LPC와의 spectral distortion (SD) 그리고 위의 과정을 통해서 구한 LPC와의 SD를 보여주고 있다. LPC 정보는 SD가 다음과 같을 때 손실을 줄일 수 있다. 평균 SD는 약 1 dB 이내, 2-4 dB의 outlier는 2% 이내에 존재해야 하며 4 dB이상의 outlier는 거의 없어야 한다 [3]. 따라서 위의 방법으로 구한 LPC는 문제가 있다고 할 수 있다. 4장에서 이 문제에 대한 해결 방안을 제시할 것이다.

부호화기	평균 SD (in dB)	Outliers (in %)	
		2-4 dB	>4 dB
G.729	1.069	3.652	0.085
제안된 부호화기	1.473	14.548	0.059

표 2. Spectral distortion 비교

Subvector elements	Codebook size
(c0)	256
(c1, c2)	64
(c3, c4)	64
(c5, c6)	64
(c7, c8)	64
(c9, c10)	64
(c11, c12)	64

표 3. MFCC 양자화를 위한 SVQ의 bit 할당

MFCC는 split 벡터 양자화 (SVQ)의 방법으로 양자화된다. MFCC의 split 방법과 양자화에 소요된 할당 bit 수는 DSR standard front-end [4]를 참고하여 수행하였다. 표 3은 SVQ와 각 subvector의 할당된 bit 수를

보여준다. MFCC 양자화기는 LBG 알고리즘 [5]을 이용하였으며 training에 사용된 data는 총 205,100의 American, English, Korean NTT-AT 음성 database의 data이다 [6].

3. 음성 부호화기의 성능 평가

3장에서는 표준 음성 부호화기인 G.729, G.729E와 음성 특징 파라미터 변화 과정에서의 오차를 고려하지 않은 음성 부호화기의 음질과 음성 인식률에 대해서 서술한다.

3.1 음질 성능 평가

제안된 음성 부호화기의 음질 측정은 perceptual evaluation of speech quality (PESQ) [7]를 이용해 수행하였다. 실험에 사용한 음성 데이터는 8개의 한국어 낭독 문장 음성으로 화자는 남자 2명, 여자 2명으로 구성되었다. 각 음성 데이터는 9000 frame으로 구성되어 있으며 8 kHz로 표본화되었다.

표 4는 G.729, G.729E 그리고 음성 특징 파라미터의 변환 과정에서의 오차를 고려하지 않은 부호화기의 평균 PESQ 점수를 보여주고 있다. 보상 전 부호화기는 G.729E보다 약 0.1 MOS 낮게 측정되었으며 G.729와 비슷한 성능을 보이고 있다.

화자	G.729E (11.8kbps)	G.729 (8kbps)	보상 전 부호화기 (10.6kbps)
남자	4.147	3.968	3.959
여자	3.910	3.729	3.725
평균	4.028	3.848	3.842

표 4. PESQ 점수 비교

3.2 음성 인식률 평가

제안된 음성 부호화기에 사용된 음성 파라미터를 이용해 학습된 ARS 시스템의 인식 성능을 테스트하였다. 또한, 제안된 음성 부호화기의 인식 성능은 1) 손상 없는 음성신호로부터 얻은 MFCC를 이용한 시스템(단말기 기반 방식), 2) ETSI DSR front-end compression algorithm [4]으로 양자화된 MFCC를 이용한 시스템(단말기/서버 기반), 3) G.729과 G729E에 의해 복원된 음성 신호를 이용한 시스템(서버 기반)의 성능과 비교했다.

인식 테스트는 [8]에서 사용한 procedure를 따랐다. 표 5는 각 ASR 시스템들의 Word Error Rates (WERs)을 보여준다. 첫 번째 행과 두 번째 행에서는 단말기 기반 방식과 서버 기반 방식의 인식 성능을 각각 보여준다. 반면, 서버 기반 방식의 인식 결과는 세 번째 열에서 볼 수 있다. 단말기/서버 기반 방식은 단말기 기반 방식에 비해 WER이 3.4% 증가하였다. 여기에서 발생하는 성능 저하는 양자화에 따른 것으로 보인다. G.729와 G.729E를 사용한 시스템의 경우, 단말기 기반 방식에 비하여 WER이 각각 12.4%, 17.5%로 증가하여 현격한 인식 성능 저하를 보인다. 하지만, 제안된 음성 부호화기를 사용한 서버 기반 방식은 WER이 2% 증가하였다. 이런 결과들로부터, 제안된 음성 부호화기가 ETSI DSR 시스템과 비교할만한 인식 성능을 가지면서 음성 신호를 복원할 수 있다는 것을 알 수 있다.

ASR Configuration		Average
단말기 기반	Uncoded	13.67
단말기/서버 기반	ETSI	14.13
서버 기반	G.729E (11.8kbps)	15.36
	G.729 (8kbps)	16.06
	Proposed (10.6kbps)	13.93

표 5. Aurora database 2 (multi-condition)을 이용한 ARS 시스템들의 인식 성능 평가 (Word Error Rate, %)

4. LSF Error 보상에 의한 음질 향상

표 2에서 보듯 MFCC를 LPC로 변환하는 과정을 통해 구한 LPC는 원래 음성으로부터 직접 구한 LPC와 일치하지 않는다. MFCC를 LPC로 바꾸는 과정을 다시 살펴보면 그 원인에 대해 몇 가지 사실을 알 수 있다. 첫째, 역 DCT 과정에서 문제가 발생한다. 음성 인식에 필요한 MFCC는 총 13개이다. 따라서 23개의 MFCC 중 13개의 MFCC를 부호화한 후 전송하게 되는데 이 13개의 MFCC로부터 23개의 MFCC를 구할 때 정확한 원래 값을 복원할 수 없다. 두 번째 원인은 spectrum을 구하기 위한 과정에서 발생한다. 즉 선형 보간을 할 때 mel-filtering 이전의 값을 정확히 복원할 수 없는 것이다.

이런 문제점을 해결하기 위해 그림 4의 과정을 거쳐 파라미터 변환 과정의 오차를 보정하게 하였다. 즉 원 음성으로부터 구한 Line spectral frequency (LSF, ω)와 MFCC로부터 구한 LSF ($\tilde{\omega}$)의 차이를 부호화하여 전송하고 이를 이용하면 오차를 보정하게 되는 것($\hat{\omega}$)이다. 표 6은 보상 파라미터 양자화에 필요한 codebook index를 추가한 음성 부호화기의 bit 할당 정보이다.

Codebook은 LBG 알고리즘 [5]을 이용해 10개의 LSF를 5개씩 분리하여 SVQ의 방법으로 생성했다. 이 때 각 LSF에 G.729에서 사용한 weight factor를 사용해 양자화 오차에 따르는 LSF의 SD 저하를 최소화하였다. 두 개의 codebook은 6 bit의 entry를 갖게 되어 개선된 음성 부호화기는 12 bit이 추가되어 11.8 kbps의 전송률을 갖게 된다. Training에 사용된 DB와 조건은 MFCC 양자화의 그것과 같다.

표 7는 오차 보정 방법을 이용하였을 때의 spectral distortion을 보여주고 있다. 즉 개선된 음성 부호화기는 Paliwal이 밝힌 기준 [2]에 부합하게 되며 이는 G.729보다 좋은 성능을 나타내게 된다. 즉, G.729의 LSF 양자화에 따른 error 보다 그림 4의 과정을 통한 error가 적게 나타나는 것이다.

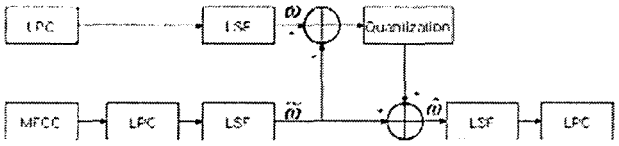

그림 4. LSF error 보상 과정

Parameter	Subframe		Frame
	1	2	
MFCC	-		44
LSF error codebook index	-		12
Adaptive codebook index	8	5	13
Pitch parity	1	-	1
Fixed codebook index	13	13	26
Fixed codebook sign	4	4	8
Codebook gain	7	7	14
Total			118

표 6. 개선된 음성 부호화기의 bit 할당 정보

부호화기	평균 SD (in dB)	Outliers (in %)	
		2-4 dB	>4 dB
G.729	1.069	3.652	0.085
제안된 부호화기	0.903	1.227	0.004

표 7. Spectral distortion 비교

표 8은 오차 보정 방법을 이용하여 구현한 개선된 11.8 kbps의 음성 부호화기의 음질 성능을 보여준다. 즉, 개선된 부호화기는 G.729보다 0.04 MOS 높게 측정되고 G.729E보다는 0.159 MOS 낮게 측정되었다.

개선된 부호화기의 음성 인식률은 개선 전 부호화기와 같은 MFCC를 사용하므로 3장에서 나타난 결과와 일치한다.

화자	G.729E (11.8kbps)	G.729 (8kbps)	개선된 부호화기 (11.8kbps)
남자	4.147	3.968	4.004
여자	3.910	3.729	3.769
평균	4.028	3.848	3.886

표 8. PESQ 점수 비교

5. 결론

본 논문에서 우리는 이동통신 환경 하에서 서버 기반의 음성 인식 시스템에 사용되는 음성 부호화기를 제안하였다. 제안된 음성 부호화기는 음성의 spectral envelope을 구하기 위해 음성 인식 파라미터인 MFCC를 사용하였으며 MFCC 양자화, MFCC로부터 LPC로의 변환, 그 변환 과정에서의 오차 보정에 대한 방안을 사용하였다. 그 결과 11.8 kbps의 전송률을 갖게 되었다. 제안된 음성 부호화기는 같은 전송률을 갖는 G.729E보다 음성 인식 능력이 뛰어나며 음질은 G.729와 G.729E 사이의 성능을 보이고 있다. 따라서 음성 부호화기로서 뛰어난 음성 인식 능력을 보유하게 되는 것이다.

참고문헌

1. H. K. Kim and R. V. Cox, "A bitstream-based front-end for wireless speech recognition on IS-136 communications system," IEEE Trans. Speech and Audio Process., vol 9, no. 5, pp. 558-568, July, 2001.
2. ITU-T Recommendation G.729, Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction (CS-ACELP), 1996.
3. K. K. Paliwal, B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," IEEE Trans. Speech and Audio Process., vol 1, no. 1, pp. 3-14, Jan. 1993.
4. ETSI Standard ES 201 108 v1.1.3, Speech processing, transmission and quality aspects; Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithm, Sept. 2003.
5. Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," IEEE Trans. Commun., vol. 28, no. 1, pp. 84-95, Jan. 1980.
6. NTT-AT, Multi-lingual speech database for telephonemetry, 1994.
7. ITU-T Recommendation P.862, Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, Feb. 2001.
8. H. G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in Proc. ASR2000, Paris, France, pp. 181-188, Sept. 2000.